DOCUMENT RESUME

ED 247 299 TM 840 480

AUTHOR Loyd, Brenda
TITLE Evaluation of Log Linear Models for Detection of Item Bias: A Comparison Across Samples.
PUB DATE Apr 84
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Entrance Examinations; Goodness of Fit; High Schools; *Mathematical Models; Minimum Competency Testing; *Racial Bias; Statistical Analysis; *Test Bias; *Test Items
IDENTIFIERS *Log Linear Models

ABSTRACT

The purpose of this study was to evaluate the use of log-linear models in detecting test item bias. The evaluation of the log-linear approach in this regard will focussed on the following: (1) the stability of the procedure across replications--the effect of sampling error across random samples of ethnic groups; (2) the amount of artifactual bias--analyzed by comparing results for different ethnic groups with the results from two random equivalent white groups and with the results from two white groups which differ in measurability; and (3) the interpretability of the items identified as deviant. Two sets of items were selected for this study: one from the American College Testing program and one from a high school minimum competency test. Although the possible effect of sample size needs to be examined further, the results suggest that methods based on log-linear models could reasonably be used to examine sets of items for potential bias. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED247299

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

EVALUATION OF LOG LINEAR MODELS FOR DETECTION OF ITEM BIAS: A COMPARISON ACROSS SAMPLES

Brenda Loyd

University of Virginia

Paper presented at the annual meeting of the
National Council on Measurement in Education

New Orleans, April 1984

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. H. Loyd

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 088

EVALUATION OF LOG LINEAR MODELS FOR DETECTION OF ITEM BIAS: A COMPARISON ACROSS SAMPLES

In recent years there has been a growing interest in exploring bias in educational tests. These tests are constructed to measure a particular achievement or ability. When individuals from different groups (culture, race, gender, etc.) with the same level of ability have different probabilities of success on a particular test, the test is said to be biased. Such bias can be defined as differential validity, and a test is considered biased if it measures different constructs for different groups or if it measures the same construct with different degrees of accuracy.

One approach to the study of test bias is to examine bias in individual test items. Many statistical methods have been developed to detect such "item bias" (see Berk, 1982). Most of these methods rely on measuring an item-by-group interaction as a potential indication of bias (Shepard, 1983), because this removes the effects of the overall difference between groups as an indication of the differential validity of an item. These methods differ in theoretical basis, statistical complexity, and cost in terms of sample size, computer use, and test developer's time. They range from the most complex techniques, employing the three-parameter item characteristic curve, to the simpler techniques such as the various chi-square methods or the transformed item difficulty method.

Although these methods are often referred to as methods which detect "item bias," it would be more accurate to describe them as methods which detect differential item performance or deviant item performance (Carlton & Marco, 1982). Without examining possible reasons why an item performs differently, bias cannot be inferred. Some items may be identified as performing differentially for two groups simply because of the particular

statistical method selected to identify item-by-group interactions; that is, the difference in the performance of the items could be due to statistical artifact.

Several studies have been conducted to compare the various methods (Ironson & Subkoviak, 1979; Rudner, Getson, & Knight, 1980; and Shepard, Camilli, & Averill, 1981). These studies examined the consistencies and inconsistencies among the various techniques. The results of these studies tend to suggest that methods based on three-parameter item response theory are performed because of the sample invariant properties of the model. Differences in means between two groups should not affect the item characteristic curves calibrated separately for two groups and then equated to the same scale. The difficulties or limitations of using three-parameter item characteristic curves include: (a) relatively large sample sizes (usually a minimum of 1000 subjects is suggested) and a large number of items (40 or more) are needed to permit adequate, stable estimates of item parameters and subjects abilities parameters; (b) the items need to be from a unidimensional test, i.e., the test items need to be consistent with the theoretical assumptions of item response theory; and (c) the method of equating the ability scales calibrated separately for two groups needs to be adequate.

Methods of item bias detection based on chi-square procedures (Scheuneman, 1979; Shepard, et al., 1981) are suggested as approximations of item response theory methods. Ability categories are established by specifying intervals on a total test score scale. Group membership is a second classification variable, and correct or incorrect response is the third classification variable. Within each ability interval, the probability of students answering an item correctly or incorrectly should

be the same regardless of group membership. The expected proportions are compared with the observed proportions by calculating a chi-square goodness-of-fit statistic, and the magnitude of the statistic is evaluated as an indication of potential item bias.

One of the difficulties in the area of "item bias" research has been the interpretation of findings. Often after a set of items have been identified as deviant, the efforts to find common factors among the deviant items, to find differences between the items identified as potentially biased and those not so identified, or to have independent judges validate the selection of the set of items have not been successful (Plake, 1980). Shepard (1983) carried out a systematic investigation of statistical artifacts in methods employing three-parameter item characteristic curves in order to assess the amount of artifactual or spurious results. In an effort to distinguish the artifactual from the more substantial and thus clarify interpretation of item bias research results, Shepard

1. used cross-validation or replications with random samples of white and black samples,
2. compared two random equivalent white groups to serve as a neutral baseline, and
3. compared two groups of white examinees which differed in mean ability representing the average differences expected between the two groups, and allowing for a pseudo-ethnic comparison as an additional baseline from which to compare results.

The purpose of this study is to evaluate the use of log-linear models in detecting "item bias." The evaluation of the log-linear approach in this regard will focus on the following:

a. the stability of the procedure across replications (effect of sampling error across random samples of ethnic groups);

b. the amount of artifactual bias (to be analyzed by comparing results for different ethnic groups with the results from two random equivalent white groups and with the results from two white groups which differ in measurability); and

c. the interpretability of the items identified as deviant.

The log-linear model

The log-linear model is used in the analysis of categorized data. When the variables of interest are categorized, the data is usually presented in the form of a contingency table, which is a table of counts. The sources of variance in a contingency table can be measured by examining cross-product ratios of the counts in the tables. Models can be constructed by using logarithms in transformations of the cross-product ratios.

A general approach to the analysis of log-linear models (Haberman, 1978) will involve: (1) selection of a model plausible for the data; (2) estimation of the parameters in the model from the data, usually by the method of maximum likelihood; and (3) testing of the model's accuracy from the parameter estimates, usually with Pearson or likelihood ratio chi-square tests which measure the overall compatibility of the model and the data.

1) Selection of a model. The log-linear model may best be understood by beginning with a 2x2 table of counts. The counts are determined by data being cross-classified into the levels of categorical variables. Given two ethnic groups and two response levels, correct and incorrect, we could generate a contingency table

	Response		
	X_{11}	X_{12}	X_{1+}
	X_{21}	X_{22}	X_{2+}
	X_{+1}	X_{+2}	N

	Response		
	P_{11}	P_{12}	P_{1+}
	P_{21}	P_{22}	P_{2+}
	P_{+1}	P_{+2}	1

Figure 1

as indicated in Figure 1 where X_{11}, X_{12}, X_{21}, X_{22} are frequencies or counts. The variability among cell counts would be due to the effect of the row variable (ethnic group), the column variable (response), any systematic relationship between the two variables, and chance or error. We assume that the counts are a sample from the population of interest, and we can describe a population contingency table by the cell probabilities (see Figure 1).

If group membership and item response are independent, then the conditional probability of getting an item count would be the same for group one and group two; i.e.,

$$\frac{P_{11}}{P_{1+}} = \frac{P_{21}}{P_{2+}}$$

However, to measure the amount of dependence we consider the ratio of the odds of getting an item correct for group one and group two:

$$\frac{\dfrac{P_{11}}{P_{12}}}{\dfrac{P_{21}}{P_{22}}} = \frac{P_{11}\ P_{22}}{P_{21}\ P_{12}}$$

The logarithm of the odds ratio or cross-product ratio would be:

$$\log \frac{P_{11}\ P_{22}}{P_{21}\ P_{12}} = (\log P_{11} + \log P_{22}) - \log P_{21} - \log P_{12}.$$

When the two variables are independent, the ratio $\frac{P_{11}\ P_{22}}{P_{21}\ P_{12}}$ equals one and the log of the ratio equals zero. Therefore, no systematic relationship would be represented by zero. When P_{11} and P_{22} are large relative to P_{21} and P_{12}, which would reflect an interaction or relationship between group membership and item response, the log of the cross product ratio will be larger than zero.

Other sources of variance can be measured by taking the logarithm of the odds ratios; e.g.,

$$\log \frac{P_{11}\ P_{21}}{P_{12}\ P_{22}} = \log P_{11} + \log P_{21} - \log P_{12} - \log P_{22}$$

represents the variance from the column variable.

When the probabilities between the two columns are consistently different in magnitude, the absolute value of the expression will increase. Therefore, the degree to which the logarithm of the cross-product ratio differs from zero reflects the degree to which the column variable influences the cell counts. This procedure can be extended so that all sources of variance in a contingency table can be described by logarithms of cross-product ratios.

In addition to describing individual sources of variance, it is possible to describe models for the tables of counts that are linear in the logarithm of the cell probabilities. A general log-linear model for a 2x2 table is

$$\log P_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)},\ i = 1, 2,\ j = 1, 2.$$

The parameter $u_{1(i)}$ represents the main effects of variable 1 (group membership), $u_{2(j)}$ represents the main effects of variable 2 (item response), and $u_{12(ij)}$ represents the interaction between the two variables.

A model which has as many parameters as there are independent cell probabilities is called a saturated model; i.e., it contains all possible sources of variance. If in the model for a 2x2 contingency table in Figure 1, we set all $u_{12(ij)}$ equal to zero and obtained a model in which the two variables are independent, this would be an unsaturated model, because the model would have fewer fitted parameters than independent cell probabilities.

The general log-linear technique can be extended to larger tables and tables of higher dimensions (see Bishop, Feinberg, & Holland, 1975; Haberman, 1978; Feinberg, 1981). Models for higher dimension tables can be expressed in sets of 2 x 2 tables. Given three variables, such as group, responses, and ability (each represented in two categories), there are several possible linear models. Table 1 presents examples of cell probabilities and their corresponding log-linear models.

In the model of complete independence among the three variables (see I, Table 1) the probability of a given cell is the joint product of the corresponding marginal probabilities. There is no interaction among variables, so the model includes only main effects.

In the log-linear model where response and ability are each independent of group membership (see II, Table 1), the relationship of cell probabilities to marginal probabilities is $P_{ijk} = P_{+jk} P_{i++}$. This log-linear model includes all three main effects and the interaction between response and ability. There are no u_{12} or u_{13} parameters represented in the model.

The conditional independence of group and response given level of ability is represented in III in Table 1. The cell probabilities are

obtained as follows: $P_{ijk} = (P_{i+k}P_{+jk})/p_{++k}$. The $u_{12(ij)}$ parameters are not represented in this log-linear model.

The model of no second-order interaction (see IV, Table 1) indicates that group and response are not independent, but the amount of dependence as measured by the cross-product ratio is the same at both levels of ability. This concept of equal dependence would hold for interactions among each pair of variables.

(2) Estimation of parameters. After selecting the most appropriate log-linear model to describe the realtionship among the cell probabilities in a contingency table, the next step is to estimate the expected cell counts and corresponding u-parameters assuming the model is true. Estimations of parameters are computed by employing the principle of maximum likelihood. Maximum likelihood estimates are functions of sufficient statistics, and for log-linear models which are "hierarchical" there is a one-to-one correspondence between the set of parameters and the set of margin totals as minimal sufficient statistics. Hierarchical models assume that whenever a u-term involving a set of variables is equal to zero and not included in the model, then all higher-order terms having that set of variables also equal zero and are not included in the model. For example, whenever the parameters $u_{12(jk)}$ are zero, in a hierarchical model the parameters $u_{123(ijk)}$ are also zero. (Non-hierarchical models are omitted from consideration in this discussion and the methods used are restricted to the class of hierarchical models.)

(2) Testing model fit. After estimates are obtained from a set of data, the next step is to test whether the model adequately describes the observed data. Two tests of goodness-of-fit are used: the likelihood-ratio statistic,

$$G^2 = 2\Sigma \text{ (observed) } \log \frac{\text{observed}}{\text{expected}}$$

and the Pearson statistic,

$$X^2 = \Sigma \frac{(\text{Observed-Expected})^2}{\text{Expected}}.$$

Each is based on the observed counts and the expected counts under the model being tested. When all estimated cell counts are non-zero, then both G^2 and X^2 are distributed as chi-square random variables with (c-p) degrees of freedom, where c is the number of cells in the table and p is the number of parameters fitted in the model.

These goodness-of-fit statistics are compared to tabled chi-square values, and models can be rejected when the probability of obtaining a chi-square value this large or larger is less than a specified probability (significance level). If the model is not rejected, then it is retained as a model which could explain the variations in the table of counts.

The likelihood-ratio statistic G^2 has an advantage over X^2 because it can be partitioned to test nested hypotheses. For example, in a three-dimensional table let the first model be

$$\log P = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23},$$

which is the model of no second order interaction. Let the second model be

$$\log P = u + u_1 + u_2 + u_3 + u_{13} + u_{23},$$

the model of conditional independence. Model two is a special case of model one since all the parameters in model two are also in model one. If g^2 and G^2_2 are the likelihood-ratio statistics for testing models one and two, and df and df_2 are the degrees of freedom associated with the tests respectively, then $G^2_{2\,1} = G^2_2 - G^2_1$ will be approximately distributed as a chi-square statistic with $df_2 - df_1$ degrees of freedom and can be used to address the question, "Is model two tenable given that we believe model one

is tenable?" A calculated value ($G^2_2 - G^2_1$) which does not exceed the critical value of the table chi-square would indicate that the second model (the model of conditional independence) would be tenable, and we would choose model two over model one. A value which exceeds the critical value would indicate that the difference between the two models was not due simply to random variation, and we would then choose model one over model two.

Bias detection with log-linear models.

Two approaches to the use of log-linear analysis for the detection of potential bias in items will be presented. The general use of log-linear models is in no sense restricted to these two approaches, and given a different situation or hypothesis, a different model and procedure would be appropriate. To examine a set of items for item bias we assume that we have two or more levels of group membership as one variable, two levels of response (correct and incorrect) as a second variable, and two or more levels of ability as a third variable. Consistent with the logical basis of other methods of examining potential item bias, an item would be biased if persons of the same ability did not have the same probability of success (and the same probability of failure) on a given item. Therefore an item with a significant group-by-response interaction would flag the item as one to be analyzed for potential bias. To test for the significance of this interaction in a table of counts, the model would be

$$\log(p) = u + u_1 + u_2 + u_3 + u_{12} + u_{23}.$$

If the model of conditional independence of group and item response given ability is tenable, no bias would be indicated. The model indicates that there is no interaction between group and response given ability level, and also that there is no second-order interaction and no systematic

relationship between group, response, and ability level (i.e., $u_{13} = u_{123} = 0$). A statistically significant goodness-of-fit statistic would indicate that one or both of these terms cannot be equal to zero, and thus indicate potential bias in the item. A second approach tests the model of no second-order interaction first (i.e., $u_{123} = 0$). If that model is tenable, then the model of conditional independence of group and item response given ability ($u_{13} = 0$) will be tested as a nested hypothesis. A statistically significant goodness-of-fit statistic for either test would identify the item as potentially biased. It is possible for the second approach to flag items not identified in the first procedure.

METHODOLOGY

Items

Two sets of items were selected for study. The first set contained 40 items administered as part of the American College Testing (ACT) program. The items measure math applications and math concepts usually taught in high school curricula, including questions on basic arithmetic operations, algebra, geometry, and advanced algebra. The items were administered to high school juniors and seniors as part of the national administration. Most students who take the ACT battery are preparing for entrance into college; the test functions as a college entrance exam.

The second set of items consisted of 55 items administered as part of a high school minimum competency program for a southeastern state. This set of items measured basic arithmetic operations, reading graphs, figuring mileage on maps, knowledge of units of measurement, ability to figure tax on purchases, ability to fill out a check, among other math-related skills. These items were among the more than 200 items administered to all high school students planning to complete high school.

Samples

For each of the two sets of items, a sample of high school students was selected. For the college entrance exam, a total sample of 4163 (918 black and 3245 white examinees) was selected from students at test sites testing 10 or more black and white examinees. This was done to try to reduce differences between race groups that might be due to differences in geographic location. In order to evaluate the consistency of results over random samples, the sample was divided into two randomly equivalent subsamples (the second will be indicated in each case below by the addition of an apostrophe (') to the notation). From each subsample (I and I'), four types of samples were created. The first (A and A') consisted of approximately 450 black examinees and approximately 450 white examinees. The second (B and B') consisted of 200 black examinees and 200 white examinees. The third (C and C') consisted of two randomly equivalent white groups of 200 examinees. The fourth sample (D and D') consisted of a sample of approximately 450 white examinees selected to match the ability of the black examinees in samples A and A' respectively (ability as measured by total score on the 40 items) and a random sample of 450 white examinees. This fourth sample was created to represent comparable differences between the distributions of total scores on the 40 items for the black and white groups, sometimes referred to as pseudo-ethnic groups.

For the high school minimum competency set of items, a total sample of 4000 was selected, 2000 black examinees and 2000 white examinees. From this larger sample two randomly equivalent subsamples were created (the second will be distinguished from the first by '). From each subsample, (II and II'), three types of samples were established. The first (AA and AA') consisted of 1000 black examinees and 1000 white examinees. The

second (BB and BB') consisted of 200 black examinees and 200 white examinees (each randomly selected from the larger samples). The third (CC and CC') consisted of two randomly equivalent groups of 200 white examinees.

Procedures

The log-linear models considered in this study were based on three-dimensional tables of counts. The three dimensions or variables were groups (two levels), response (two levels, correct and incorrect), and ability (three levels, low, middle, high). For the college entrance exam, ability groups were established by specifying intervals on the total score scale. A score of 13 or less was low, 14-19 was middle, and above 20 or above was high (possible scores range from 0 to 40). These levels were the same for all analyses using the college entrance exam. For the high school competency exam, a score of 44 or less was low ability, 45-49 was middle ability and 50 or above was high ability (scores could range from 0 to 55). These intervals were the same for all analyses using the high school competency exam. As may be suggested by the intervals, the two sets of items differed considerably in average difficulty for their respective samples. For the college entrance exam items, the average item difficulty was 0.49, while for the high school competency exam items, the average item difficulty was .83.

Two log-linear models were specified: the model with no second order interaction; and the model of conditional independence of group and response given level of ability (as described above). The fit of each model was calculated with the aid of the BMP computer program package. Goodness-of-fit likelihood-ratio statistics (G^2) were determined for each of the two models for each of 14 tables of counts (A, B, C, D, A', B', C',

and D' for the college entrance exam items; and AA, BB, CC, AA', BB', and CC' for the high school competency exam items).

The statistics of interest were the likelihood-ratio statistic for the model of conditional independence of group and response given level of ability, and the statistic created by taking the difference between the two models' likelihood-ratio statistics. The first statistic is distributed as a chi-square with three degrees of freedom when the model fits the table of counts. The second statistic is distributed as a chi-square with one degree of freedom under the hypothesis that the two models differ only by chance.

To identify items as potentially biased (i.e., to flag deviant items), a significance level of .005 was set for each individual comparison. This level was selected over a larger value in an effort to limit the number of items that might be flagged due to Type I error. Considering each of the eight analyses of the set of 40 items as a family of comparisons, the level of significance for the set of items would be less than or equal to .20 in each analysis. Similarly, for each of the six analyses of the set of 55 items, the significance level for the set would be less than or equal to 0.27. Items were identified as potentially biased if the model of conditional independence was rejected (method 1) or if the two models differed by more than chance (method 2).

Pearson correlations were computed among the statistics for each sample across a set of items. Kappa coefficients were also calculated to assess agreement among items identified as potentially biased for each of the two methods.

The results of the above analyses were then examined in terms of stability of results, statistical artifact, and interprebility. To address the first, the A-A', B-B', AA-AA', BB-BB' comparisons were examined. These

comparisons indicated how results replicate across random samples and compare across two sample sizes.

The amount of statistical artifact in the results was examined in two ways. First, B and B' results were compared with C and C' results respectively. This permitted the evaluation of the results of black-white comparisons in relation to a baseline of differential responses of two randomly equivalent groups. Next, A and A' results were compared with D and D' results respectively. This allowed comparison of black-white differences with white-white differences when the two white groups differ in mean ability.

Finally, the interprebility of the results was explored by examining items identified as potentially biased for consistencies or inconsistencies in order to evaluate the practical implications of using the log-linear model for detection of biased items.

RESULTS AND DISCUSSION

In Table 3 are two tables of counts and the results of the log-linear analyses for two items from the college entrance exam. The analysis for item 29 indicated that the model of conditional independence of group and response for a level of ability could reasonably fit the table of counts. The analysis for item 3 indicated that the model of conditional independence did not adequately fit the data; this model was therefore rejected in favor of a model which included a group-by-response interaction term.

Summaries of the two statistics are presented in Tables 4 and 5. Table 4 presents the statistics for each of the two methods averaged over the items for each sample. For the first set of items, the means of the fit statistics for the conditional model ranged from 2.75 to 5.98. The

means for the differences between the fit statistics of the two models ranged from 1.03 to 3.18. For the second set of items, the means of the fit statistics for the conditional model ranged from 3.24 to 9.17 and the means for the differences between fit statistics ranged from 1.56 to 6.46. If the model of conditional independence is correct for all items, the expected value would be three. Therefore, these mean values suggest that the model of conditional independence fits the first set of items better than the second set.

Table 5 presents Pearson correlation coefficients for the statistics based on the two methods for each comparison of sample pairs. The coefficients for the comparisons of A with A', B with B', AA with AA',, and BB with BB' indicate the consistency in the magnitude of the fit statistics for the items across random samples. The correlation coefficients range from .26 to .86 and suggest that there is some consistency across random samples. The second method based on difference between fit statistics has consistently higher correlations for these four sets of comparisons, ranging from .41 to .91.

The differences among results for different sample sizes were not consistent for the two sets of items. For the first set of items, the correlations between statistics based on totals of 400 and 900 were not consistently or dramatically different. However, there was a large difference between correlations of statistics based on totals of 400 and 2000 (with correlations of .30 and .41 versus .86 and .91 respectively).

To address the question of the amount of chance agreement represented in the correlations, the consistency of the results for two black-white analyses (as represented in B-B' and BB-B' correlations) can be compared with analyses based on white-white comparisons (C-C' and CC-CC'). The

correlations for the black-white analyses range from .30 to .47. The corelations for the white-white analyses range from -.11 to .20. This question can also be addressed by considering the correlations between a black-white analysis and a white-white analysis. As indicated by the correlations in the last two rows of Table 5, the correlations between these analyses seem to fluctuate about zero, ranging from -.11 to .06. These results would suggest that the correlations established between black-white analyses are greater than would be expected from analyses based on randomly equivalent white groups.

To pursue further the comparison of black-white analyses with white-white analyses, the correlations for the A-A' analyses can be compared to those for the D-D' analyses. This comparison (although based on limited evidence) would suggest that there is a difference between the black-white analyses and the white-white (pseudo-black) analyses. However, the results might also suggest that the log-linear model fit statistics are influenced by differences in mean ability as well as group-by-response interaction.

The correlations summarizing the consistency of the fit statistics across samples provide basic support for the use of these two methods based on log-linear models. For cmparisons of group results where bias is a possibility (as with the black-white analyses), the methods produce greater consistency than when there is no bias expected (as in the white-white comparisons).

In addition to the magnitude of the statistics, it is important to evaluate the consistency with which items are classified as potentially biased or nonbiased. Table 6 presents the number of items identified as "biased" and "unbiased" for each set of items for each analysis. The agreement across random samples is indicated by the kappa coefficients. For each set of items, the black-white analyses identified some of the same

items as biased. The white-white analyses did not identify the same items as biased. For the set of 40 items, only one item was consistently identified as deviant across samples for analyses, A and A'. For the set of 55 items, seven common items were identified across the two analyses, AA and AA'. The pattern of items identified as deviant is similar to the pattern suggested by the correlations. Items were consistently identified as deviant for analyses where deviant items could be expected (black-white analyses), while for white-white analyses no items were consistently identified.

One of the reasons for replicating analyses across samples is to see if after controlling for statistical artifact, the pattern of items which were identified as "biased" can be interpreted. Therefore, items which were consistently identified as biased were analyzed for content and format to see whether any pattern could be discerned which would indicate possible reasons for the item to perform differently for different race groups. For the set of 40 items from the college entrance exam, the one item consistently identified as biased across samples was an item measuring a geometry concept, specifically related to bisecting line segments. A second item which was consistently identified by the second method was an item measuring the concept of set theory. Both items were in a multiple-choice format which was identical to the format for the other 38 items. It is difficult to establish a pattern based on two items and difficult to interpret why an item related to geometry or algebra would be shown to perform differently for different race groups when the seven other items measuring similar content are not shown to be performing differentially. (One might assume that if the significance level used for making decisions had not been as stringent, more items might have been identified. To

examine this possibility, a .05 significance level was used for the A-A' comparison. A larger number of items were thus identified as potentially biased for each analysis individually, but none of them were consistent across the two analyses; this suggests in part that a higher significance level would have lowered the agreement across samples.)

Examining the nine items from the high school competency exam which were consistently identified as performing differentially for the two race groups, the multiple-choice format is similar across all items. The content includes two items on reading graphs, two items on comparing unit prices of goods, one item on the concept of parallelism, one item on how to fill out a check, one item measuring time, one item on metric units of measure, and one item on percentages. Again, no clear pattern emerges and the set of nonidentified items contains items similar in content that were not performing differentially by race groups.

Although the interpretation of the items identified as potentially biased was not enlightening in this particular case, the general results of this study support the usefulness of log-linear models in detecting potentially biased items. The patterns of correlations and numbers of items identified suggest that the methods used in this study identified items in samples where bias could be present and failed to identify items in samples where bias could not reasonably be expected.

The patterns found in this study are similar to those reported by Shepard (1983) for methods based on the use of three-parameter item characteristic curves (ICCs). Although this study could not provide a direct comparison between the log-linear model and three-parameter ICC methods, Ironson and Subkoviak (1979) and Shepard, et al. (1980) compared one method based on three-parameter ICCs and a chi-square procedure (referred to as the full chi-square procedure; see Ironson, 1982). The

correlations they found were .59 and .46 respectively. The procedure used in this study to determine the index of fit of conditional independence is the same as the goodness-of-fit statistic for the full chi-square procedure; therefore, we could assume that the correlations would be similar here.

In summary, this study suggests that methods based on log-linear models could reasonably be used to examine sets of items for potential item bias. These methods have the advantage of being simpler to use than models based on three-parameter ICCs in terms of availability and complexity of computer programs, and in terms of the number of items and subjects required for adequate parameter estimation (although the results of this study suggest that the possible effect of sample size for the log-linear model needs to be examined further).

REFERENCES

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press.

Bishop, Y.M.M., Feinberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge: The MIT Press.

Carlton, S. T., & Marco, G. L. (1982). Educational testing service. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 278-313). Baltimore: The Johns Hopkins University Press.

Feinberg, S. E. (1980). The analysis of cross-classified categorical data (2nd ed.). Cambridge: The MIT Press.

Haberman, S. J. (1978). Analysis of qualitative data. New York: Academic Press.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore: The Johns Hopkins University Press.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Rudner, L. M., Geston, P. R., & Knight, D. L. (1980). Biased item detection techniques. Journal of Educational Statistics, 5, 213-233.

Scheuneman, J. D. (1979). A new method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L., Camilli, G., & Averill, M. (1980, April). Comparison of six procedures for detecting item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Shepard, L., Camilli, G., & Williams, D. M. (1983, April). Accounting for statistical artifacts in item bias research. Paper presented at the annual meeting of the American Educational Research Association, Quebec: Montreal. (ERIC Document Reproduction Service No. ED 233 085).

Table 1

Selected Log Linear Models for Three Dimensional Contingency Table

Ability

Group	A_1 Response R_1	A_1 Response R_2	A_2 Response R_1	A_2 Response R_2
G_1	P_{111}	P_{121}	P_{112}	P_{122}
G_2	P_{211}	P_{221}	P_{212}	P_{222}

I.

	A_1 R_1	A_1 R_2	A_2 R_1	A_2 R_2
G_1	.192	.288	.128	.192
G_2	.048	.072	.132	.048

Model of Complete Independence of Group, Response, Ability

$$\log P_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

II.

	A_1 R_1	A_1 R_2	A_2 R_1	A_2 R_2
G_1	.18	.3	.06	.06
G_2	.12	.2	.04	.04

Model of Response by Ability Interaction and Independence of Group

$$\log (pijk) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$$

Table 1 (continued)

III.

	A_1	
	R_1	R_2
G_1	.108	.072
G_2	.252	.168

	A_2	
	R_1	R_2
G_1	.025	.075
G_2	.075	.225

Model of Conditional Independence
of Group and Response Given Ability

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}$$

IV.

	A_1	
	R_1	R_2
G_1	.2	.1
G_2	.2	.05

	A_2	
	R_1	R_2
G_1	.09	.09
G_2	.18	.09

Model of No Second Order Interaction

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

Table 2

Descriptions of Sample Size and Average Ability

Sample I	College Entrance Exam Items (40 items)		I'		
	N	X Ability		N	X Ability
A	456 black	13.9	A'	459 black	13.6
	460 white	18.9		449 white	19.1
B	200 black	13.5	B'	200 black	13.7
	200 white	19.0		200 white	19.4
C	200 white	18.7	C'	200 white	18.9
	200 white	19.2		200 white	19.4
	(randomly equivalent)			(randomly equivalent)	
D	455 white	14.4	D'	459 white	14.1
	463 white	19.1		465 white	18.4
	(pseudo-ethnic)			(pseudo-ethnic)	
Sample II	High School Competency Items (55 items)		II'		
AA	1000 black	43.4	AA'	1000 black	43.2
	1000 white	46.6		1000 white	46.7
BB	200 black	43.2	BB'	200 black	43.3
	200 white	46.5		200 white	46.8
CC	200 white	47.2	CC'	200 white	47.4
	200 white	46.3		200 white	47.5
	(randomly equivalent)			(randomly equivalent)	

Table 3
Examples of Computer Output
for Log-linear Models

USING LEVEL 3.00000 OF VARIABLE 2 QUEST
****

RACE	RESPONSE	ABILITY LOW	MEDIUM	HIGH	TOTAL
BLACK	WRONG	160	51	14	225
	RIGHT	93	88	53	234
	TOTAL	253	139	67	459
WHITE	WRONG	31	41	33	105
	RIGHT	46	93	205	344
	TOTAL	77	134	238	449

Possible Bias Indicated

TOTAL OF THE OBSERVED FREQUENCY TABLE IS 908

MODEL	DF	LIKELIHOOD-RATIO CHISQ	PROB.	PEARSON CHISQ	PROB.
Conditional Independence A, G, R, AG, AR	3	15.67	.0013	15.90	.0012
No Second Order Interaction A, G, R, AG, AR, GR	2	3.29	.1931	3.29	.1934

***** OBSERVED FREQUENCY TABLE 1

USING LEVEL 29.0000 OF VARIABLE 2 QUEST
****

RACE	RESPONSE	ABILITY LOW	MEDIUM	HIGH	TOTAL
BLACK	WRONG	192	94	33	319
	RIGHT	61	45	34	140
	TOTAL	253	139	67	459
WHITE	WRONG	56	92	114	262
	RIGHT	21	42	124	187
	TOTAL	77	134	238	449

No Bias Indicated

TOTAL OF THE OBSERVED FREQUENCY TABLE IS 908

MODEL	DF	LIKELIHOOD-RATIO CHISQ	PROB.	PEARSON CHISQ	PROB.
Conditional Independence A, G, R, AG, AR	3	.38	.9437	.39	.9427
No Second Order Interaction A, G, R, AG, AR, GR	2	.29	.8635	.29	.8629

Table 4

Means of Fit Statistics Across Items

Sample	1	2	Sample	1	2
A	4.18	1.79	AA	9.17	6.46
A'	5.98	2.45	AA'	9.05	5.10
B	4.21	2.18	BB	3.93	2.13
B'	4.45	2.41	BB'	4.13	1.87
C	2.75	1.03	CC	5.84	2.50
C'	5.05	3.18	CC'	3.28	1.56
D	5.01	1.74			
D'	3.87	1.20			

1) Likelihood ratio statistic for model of conditional independence

2) Difference between likelihood ratio statistic for the model of conditional independence and the model of no second order interaction.

Table 5

Correlations of Fit Statistics

Samples	Statistic 1	Statistic 2	Samples	Statistic 1	Statistic 2
A with A' (Black/white vs. black/white)	.26	.51	AA with AA'	.86	.91
B with B' (Black/white vs. black/white)	.33	.47	BB with BB'	.30	.41
C with C' (White/white vs. white/white)	-.11	-.09	CC with CC'	.20	.01
D with D' (White/white [pseudo-black] vs. white/white [pseudo-black])	.08	.17			
B with C' (Black/white vs. white/white)	-.04	-.11	BB with CC'	.03	-.02
C with B' (white/white vs. black/white)	.04	-.06	CC with BB'	.06	.05

Table 6

Number of Items Identified as Biased
Kappa Coefficients and Percent Agreement Across Samples

Method 1

A

		NB	B	
A'	NB	35	1	36
	B	3	1	4
		37	2	

Kappa = .29
% agreement = 90

B

		NB	B	
B'	NB	38	0	38
	B	2	0	2
		40	0	

Kappa = 0
% agreement = 95

C

		NB	B	
C'	NB	37	0	37
	B	3	0	3
		40	0	

Kappa = 0
% of agreement = 93

D'

		NB	B	
D	NB	36	1	37
	B	3	0	3
		39	1	

Kappa = -.04
% of agreement = 90

Method 2

A

		NB	B	
A'	NB	32	1	33
	B	5	2	7
		37	3	

Kappa = .22
% agreement = 85

B

		NB	B	
B'	NB	36	1	37
	B	2	1	3
		38	2	

Kappa = .35
% agreement = 93

C

		NB	B	
C'	NB	36	0	36
	B	4	0	4
		40	0	

Kappa = 0
% of agreement = 90

D

		NB	B	
D'	NB	35	1	36
	B	4	0	4
		39	1	

Kappa = -.04
% of agreement = 88

Table 6 (continued)

Method 1

AA

		NB	B	
AA'	NB	42	3	45
	B	3	7	10
		45	10	

Kappa = .63
% agreement = 89

BB

		NB	B	
BB'	NB	49	3	52
	B	3	0	3
		52	3	

Kappa = -.05
% agreement = 89

CC

		NB	B	
CC'	NB	53	2	55
	B	0	0	0
		53	2	

Kappa = 0
% agreement = 96

Method 2

AA

		NB	B	
AA'	NB	42	2	44
	B	2	9	11
		44	11	

Kappa = .78
% agreement = 93

BB

		NB	B	
BB'	NB	47	4	51
	B	3	1	4
		50	5	

Kappa = .15
% agreement = 8[illegible]

CC

		NB	B	
CC'	NB	49	5	54
	B	1	0	1
		50	5	

Kappa = -.02
% agreement = 92